Environmental Health Perspectives
Vol. 32, pp. 143-156, 1979

Designing Case-Control Studies

by Takashi Yanagawa*

Identification of confounding factors, evaluation of their influence on cause-effect associations, and the introduction of appropriate ways to account for these factors are important considerations in designing case-control studies. This paper presents designs useful for these purposes, after first providing a statistical definition of a confounding factor. Differences in the ability to identify and evaluate confounding factors and estimate disease risk between designs employing stratification (matching) and designs randomly sampling cases and controls are noted. Linear logistic models for the analysis of data from such designs are described and are shown to liberalize design requirements and to increase relative risk estimation efficiency. The methods are applied to data from a multiple factor investigation of lung cancer patients and controls.

Introduction

Case-control studies play an essential role in studying cause-effect relationships in human populations (*1-3*). Applications of these studies are becoming more and more complex, as was pointed out by McKinlay (*4*) in her recent review, with emphasis increasingly being given to the investigation and estimation of multivariate sources of variation. Thus modern multivariate statistical techniques could and should be applied in both the design and analysis of such case-control studies. This requires that statisticians understand many important ideas traditionally developed in epidemiology and that epidemiologists obtain a knowledge of complicated multivariate statistical techniques. It is hoped that this paper, written by a mathematical statistician beginning the study of epidemiology, may aid epidemiologists and statisticians in their mutual understanding.

The paper reviews recent developments in the design of case-control studies, including confounding, overmatching, and effect modification from a theoretical viewpoint after introducing a statistical definition of a confounding factor. Methods of identification of confounding factors, evaluation of their influence on the measurement of cause-effect associations, and a method to control for their influence are discussed. Linear logistic models to aid in this process are introduced and applied to the analysis of a set of data from lung cancer patients and controls.

*Department of Mathematics, Faculty of Science, Kyushu University 33, Fukuoka 812, Japan.

Case-Control Studies

Let us consider the exposure and disease association in the population. Table 1 provides an example of the distribution of a rare disease and exposure to a single substance in the population; the prevalence rate of disease is 55/100,000, and half the population is exposed to the factor.

If the marginal column totals are fixed, then we have the cell probabilities given in Table 2. Table 2 suggests that if equal numbers of exposed and unexposed individuals were to be followed, well over 10,000 unexposed persons would be required before cases of disease could be expected. This type of

Table 1. Population distribution of exposure to a factor and disease status.

	Unexposed	Exposed	Total
Disease	5	50	55
Disease-free	49,995	49,950	99,945
Total	50,000	50,000	100,000

Table 2. Probabilities of the disease in Table 1 when the marginal of the exposure is fixed.

	Unexposed	Exposed
Disease	0.0001	0.0010
Disease-free	0.9999	0.9990
Total	1	1

Table 3. Probabilities of the exposure in Table 1 when the marginal of the disease is fixed.

	Unexposed	Exposed	Total
Disease	0.1	0.9	1
Disease-free	0.5	0.5	1

Table 4A. Joint distributions of exposure to a factor E and a variable Z in cases and controls.

	z_1		z_2		
	$\bar{E}$	E	$\bar{E}$	E	Total
D	0.032	0.162	0.714	0.092	1
$\bar{D}$	0.234	0.234	0.520	0.012	1

follow-up study is called a prospective or cohort study.

On the other hand, if the row marginal totals are fixed, then we have the cell probabilities given in Table 3. These numbers suggest that under 100 diseased and disease-free individuals would be required. Such a study is called a retrospective, or case-control study, since past exposure to the factor is determined retrospectively among diseased and disease-free individuals. MacMahon and Pugh (*3*) have discussed several reasons for their preference of the terms "cohort" and "case-control" over the terms "prospective" and "retrospective." We shall follow their preference throughout this paper.

Case-control studies may be, as was shown in the above example and had been pointed out by Mantel and Haenszel (*1*), the only feasible approach to the study of cause-effect association for especially rare diseases, since a cohort study may prove too expensive to consider, and the study size required to obtain a respectable number of cases completely unmanageable.

Both case-control studies and cohort studies are able to study only cause-effect association, not prove cause-effect relationships. Mantel and Haenszel (*1*) have warned that "the findings of a retrospective study are necessarily in the form of statements about association between diseases and factors, rather than about cause and effect relationships." Such studies play an important role in the chain of scientific investigation of suspected cause-effect relationships. They are a part of the cyclic process of formulating hypotheses, examining the hypotheses against existing data, and then (testing) the hypotheses through various epidemiologic and experimental studies. The most significant purpose of epidemiology is the prevention of disease. For that purpose it may not be necessary to identify the causal factors precisely.

Recognition of a cause-effect association, which is sometimes called epidemiologic association, can play an essential role in the prevention of disease. MacMahon and Pugh (*3*) made this point as follows: "The evaluation of the causal nature of a relationship, in the absence of direct experiment, is neither easy nor objective. Differences of opinion resulting from the subjective assembly and interpretation of evidence are common. Caution in judging relationships to be causal is laudable. On occasion, however, such caution appears to be carried to an unrealistic extreme. When the derivation of experimental evidence is either impracticable or unethical, there comes a point in the accumulation of evidence when it is more prudent to act on the basis that the association is causal rather than to await further evidence. If there is controversy or argument, it should center around the decision as to where this point lies, and not on the unanswerable question of whether the causal hypothesis is not proven."

When marked increases in disease frequency in a short period of time are observed, sudden exposure to a single factor can generally be suspected, and it would not be difficult to elucidate the cause-effect association by case-control studies. Applications of such studies to the more difficult problems of cancer epidemiology were begun in 1950, and the usefulness of this approach was established in the much-publicized studies clarifying the smoking and lung cancer relationship. Since the publication of the milestone paper by Mantel and Haenszel (*1*) which provided a methodology for the design and analysis of modern case-control study, studies have been undertaken to examine cause-effect associations with cancers of almost all sites.

Application of case-control studies to cancer epidemiology requires careful attention in the design of such studies, since effects of confounding variables such as sex and age, measurement errors, selection of controls, etc., could exaggerate or mask the association. One limitation of the case-control study is that it often depends upon information retrieved from the memories of individuals, or from poorly written documents. Because of these problems, case-control studies are often considered to be inferior to cohort studies. But where cancer epidemiology is concerned, this may not be true. Such problems may just be common features of studying human populations. Even if we could devise randomization or stratification in cohort studies, we would have no choice but to await the onset of disease. If the disease had a long latency period, follow-up could prove to be difficult or im-

possible, and we could expect to face problems similar to those generated by case-control studies. Problems associated with case-control studies have been discussed by many authors, including Mantel and Haenszel (*1*), Cochran (*5*), MacMahon and Pugh (*3*), Lilienfeld (*6*), and others. A review and an extensive list of papers on the design and analysis of observational studies have been published by McKinlay (*4*).

In the following sections we shall use the tools of theoretical statistics to examine various ideas which were introduced mainly by epidemiologists; emphasis will be placed on confounding, effect modification, and the logistic linear model, all of which are important in the design and analysis of case-control studies.

A Measure of Association

We shall introduce a measure of association between an exposure and the disease we wish to study. Since a primary goal of a case-control study is to reach the same conclusion as would have been obtained from a cohort study, if one had been done under complete control, we choose to define the measure within a prospective framework. Let $P\ (D/E)$ [$P\ (D/\bar{E})$] be the probability of disease in an individual previously exposed (unexposed) to a factor, $P\ (\bar{D}/\bar{E})$ [$P\ (\bar{D}/E)$] be the probability of being disease-free for an individual previously exposed (unexposed) to the factor. The relative risk RR of disease due to the factor is defined by Eq. (1):

$$\mathrm{RR} = \frac{P\ (D/E)}{P\ (D/\bar{E})} \tag{1}$$

Cornfield (*7*) showed that if the prevalence of the disease is small enough, the relative risk can be approximated by the odds ratio (2)

$$\psi = \frac{P(D/E)\ P(\bar{D}/\bar{E})}{P(\bar{D}/E)\ P(D/\bar{E})} \tag{2}$$

It follows from Bayes' theorem that ψ can be rewritten as in Eq. (3):

$$\psi = \frac{P(E/D)\ P(\bar{E}/\bar{D})}{P(\bar{E}/D)\ P(E/\bar{D})} \tag{3}$$

where $P\ (E/D)$ [$P\ (\bar{E}/D)$] is the probability of exposure (no exposure) among diseased individuals and $P\ (E/\bar{D})$ [$P\ (\bar{E}/\bar{D})$] is the probability of exposure (no exposure) among disease-free individuals. This representation shows that ψ may be estimated by a case-control study. ψ provides, therefore, a rationale for replacing an idealized cohort study with a case-control framework. Berkson (*8*) pointed out that the relative risk measure has several drawbacks. However, the other measures do not have the invariance property of ψ, or its function, and require outside knowledge which is frequently unobtainable from a case-control study. This and other problems of measures of association are discussed in Fleiss (*9*).

Confounding Factors

It is well known that exposure and disease association such as that between smoking and lung cancer are often influenced by such factors as sex, age, ethnic group, and others. Epidemiologists often term them *confounding factors*. The influence of confounding factors must be eliminated, either through procedures for selecting controls — by matching the controls with respect to the relevant factors — or in the analysis. However, neither an explicit definition of confounding factor nor a definitive method of evaluating its influence upon exposure and disease association has been given. In fact, which factors among many should be selected for case-control matching in studying exposure and disease association remains one of the most confusing and troublesome problems in the design of case-control studies. For example, matching on those factors known or strongly suspected to be related to disease occurrence was suggested by Mantel and Haenszel (*1*) and Worcester (*10*), among many others, whereas Miettinen (*11*) suggested matching on factors related to both exposure and disease. Hardy and White (*12*) emphasized matching factors related to exposure, although they generally agreed with Miettinen. Care must be taken in using this terminology. As was pointed out by Fisher and Patil (*13*), the phrases "related to disease" and "related to exposure", as used in the Miettinen article are ambiguous and can be understood in several different ways. To resolve this difficulty, we shall give a statistical definition of "confounding factor" and consider its relation to "relatedness."

Let z be a third variable. Assume for simplicity that z is a dichotomous variable (such as sex) taking on two values, z_1 (male) and z_2 (female). Let $P(D/E,z)$, $P(\bar{D}/E,z)$, $P(D/\bar{E},z)$, and $P(\bar{D}/\bar{E},z)$ be the probabilities of being diseased or disease-free among individuals exposed or unexposed to the factor E, as a function of z. Then $\psi\ (z)$, Eq. (4),

$$\psi(z) = \frac{P(D/E,z)\,P(\bar{D}/\bar{E},z)}{P(\bar{D}/E,z)\,P(D/\bar{E},z)} \tag{4}$$

is the odds ratio as a function of z. ψ, as given in the previous section, may be written as in Eq. (5),

$$\psi = \frac{[g(z_1)P(D/E,z_1) + g(zz)P(D/E,z_2)]\,[h(z_1)P(\bar{D}/\bar{E},z_1) + h(z_2)P(\bar{D}/\bar{E},z_2)]}{[g(z_1)P(\bar{D}/E,z_1) + g(z_2)P(\bar{D}/E,z_2)]\,[h(z_1)P(D/\bar{E},z_1) + h(z_2)P(D/\bar{E},z_2)]} \quad (5)$$

where $g(z)$ $[h(z)]$ is the distribution of z in the exposed (unexposed) population. We may take $g(z) = h(z)$ by such devices as stratification or matching, yet it is clear that ψ is influenced by the distribution of z. It is not necessary that $\psi = \psi(z_1) = \psi(z_2)$ hold. For example, let us consider the data given in Table 4. From Table 4 A-1, we have $\psi = (z_1) = \psi(z_2) = 5.06$, yet from Table 4 A-2 $\psi = 1.05$.

DEFINITION: confounding factor z is a factor which violates

$$\psi = \psi(z) \text{ for some value of } z.$$

In the above example it would be reasonable to accept $\psi(z_1) = \psi(z_2) = 5.06$ as a proper association of the exposure and the disease, and to consider $\psi = 1.05$ as an improper association biased by the confounding factor z; in other words, we may say that the influence of the confounding factor z on ψ is blocked by the stratification on z.

Stratification is applied regularly to block the influence of confounding variables. Note that matched pairs design is an extreme form of stratification, where only a case and a control are in each stratum. Generally, a 2 × 2 table is constructed for each stratum, the odds ratio is estimated and tested, and a summary statistic is calculated to summarize results obtained from all strata. Identification of confounding variables is a most difficult step in this procedure. Even if we could identify them successfully, we occasionally must ignore some factors whose influence on the association is not strong, especially if the number of cases is not large. For example, if the number of confounding variables were 10, then we would have to distribute cases among at least $2^{10} = 1024$ strata, an unfortunate situation if the number of cases were, for example, 300 or so. Therefore, in designing such a study, identification of confounding variables that exist in studying the exposure-disease relationship, evaluation of the strength of their influence, and introduction of efficient devices, such as matching, stratification, or others, to block their influence on the measure of association are essential.

Table 4A-1. Expected number of observations based on Table 4A when stratified by means of z.

	z_1[a]				z_2[a]		
	$\bar{E}$	E	Total		$\bar{E}$	E	Total
D	33	167	200	d	177	23	200
$\bar{D}$	100	100	200	$\bar{d}$	195	5	200

[a] $\psi(z_1) = 5.06$; $\psi(z_2) = 5.07$.

Table 4A-2. Expected number of observations based on Table 4A when the variable z is ignored.[a]

	$\bar{E}$	E	Total
D	149	51	200
$\bar{D}$	151	49	200

[a] $\psi = 1.05$.

Next we shall consider the work of Miettinen in relation to the term confounding factor as defined above. The terms "related" and "unrelated" are defined as follows.

DEFINITION: z is said to be related to disease when at least one of the probabilities $P(D/E,z)$ and $P(D/\bar{E},z)$ depends on z, i.e., altering the value of z changes the probability of disease among exposed or among unexposed individuals. z is said to be related to exposure when at least one of the probabilities $P(E/D,z)$ and $P(E/\bar{D},z)$ depends on z. If z is not related to disease, i.e., neither $P(D/E,z)$ nor $P(D/\bar{E},z)$ depends on z, z is said to be unrelated to disease. Similarly, if z is not related to exposure, z is said to be unrelated to exposure.

It may be proved under general conditions that $\psi(z) = \psi$ for any value of z if and only if z is unrelated to at least one of the entities exposure and disease. Therefore from our definition of a confounding factor we are led to the same conclusion as that of Miettinen: a confounding factor is one related to both exposure and disease. Although it is difficult to check whether the variable z is related to exposure in a case-control framework, it would be extremely difficult to check whether z is related to disease. Note that $P(D/E,z)$ is the absolute risk of disease due to exposure to the factor. Generally, it is impossible to study absolute risk from a case-control framework unless further information is obtained from outside knowledge.

Fortunately, however, the interpretation of "related" which will be given below makes it possible to identify a confounding factor and to evaluate its influence on a cause-effect association, even from a case-control study. Let us consider Table 5 showing the joint distribution of exposure to a factor in cases and in controls.

Table 5

	z_1		z_2		
	$\bar{E}$	E	$\bar{E}$	E	Total
D	P_{11}	P_{12}	P_{13}	P_{14}	1
$\bar{D}$	P_{01}	P_{02}	P_{03}	P_{04}	1

If we define

$$\psi(Dz/\bar{E}) = \frac{P_{11}P_{03}}{P_{01}P_{13}} \quad (6)$$

$$\psi(Dz/E) = \frac{P_{12}P_{04}}{P_{02}P_{14}} \quad (7)$$

$$\psi(Ez/D) = \frac{P_{11}P_{14}}{P_{12}P_{13}} \quad (8)$$

$$\psi(Ez/\bar{D}) = \frac{P_{01}P_{04}}{P_{02}P_{03}} \quad (9)$$

then it can be proved that the factor z is unrelated to the disease if and only if [Eq. (10)]

$$\psi(Dz/\bar{E}) = \psi(Dz/E) = 1 \quad (10)$$

and z is unrelated to the exposure if and only if

$$\psi(Ez/\bar{D}) = \psi(Ez/D) = 1 \quad (11)$$

Therefore, z is a confounding factor if and only if both of Eqs. (10) and (11) are violated. The magnitude of the violations reflects the strength of the influence of the confounding factor. The last point will be discussed further in the remaining sections. Note that the above table for the joint distribution is not stratified on z. So long as stratification and 2×2 table analysis are used in a case-control study, it is not feasible to check whether the factor is related to disease or not.

Overmatching

Miettinen (*11*) has considered another important problem: overmatching. If a factor z is unrelated to exposure, nothing is changed by matching on z. Thus matching is futile. However, if a factor z is unrelated to disease but related to exposure, matching by z decreases the efficiency (i.e., increases the variance) of estimated relative risk, although it does not change the value of the estimated relative risk itself. It can be proved that the stronger the relation to exposure, i.e., the larger the value of $\psi(Ez/\bar{D})$, the greater the decrease in efficiency. Thus in such a situation matching is harmful and should be avoided. This is the situation of overmatching discussed by Miettinen.

MacMahon and Pugh (*3*) suggested another case of overmatching: "Variables intermediate in the causal pathway between the study factor and the disease should not be matched. For example, if smoking altered blood cholesterol, which in turn was casually associated with cardiovascular disease, smoking would be considered a cause of cardiovascular disease. Yet, in a case-control study, if cases and controls are matched on cholesterol levels, no association of the disease with smoking would emerge." This suggests that, although blood cholesterol is a confounding factor, it should not be used for matching. Here we find one weakness of our statistical definition of a confounding factor. It is not feasible in the present framework to check whether the factor is intermediate in the causal pathway or not. This is essentially a point which must be resolved through medical knowledge.

As an illustration, let us consider the data summarized in Table 6. We have $\psi(z_1) = \psi(z_2) = 1.0$, whereas $\psi = 5.0$. This would be an example of overmatching if z were an intermediate factor in the causal pathway. However, if this is not the case, avoidance of matching provides a spurious association. From Table 6B we have $\psi(Dz|\bar{E}) = 405.8$,

Table 6B. Joint distributions of exposure to a factor E and a variable Z among cases and controls.

	z_1		z_2		
	$\bar{E}$	E	$\bar{E}$	E	Total
D	0.010	0.001	0.495	0.494	1
$\bar{D}$	0.746	0.074	0.091	0.089	1

Table 6B-1. Expected number of observations based on Table 6B when stratified by means of z.

	z_1				z_2		
	$\bar{E}$	E	Total		$\bar{E}$	E	Total
D	182	18	200	d	100	100	200
$\bar{D}$	182	18	200	$\bar{d}$	101	99	200

Table 6B-2. Expected number of observations based on Table 6B when the variable z is ignored.[a]

	$\bar{E}$	E	Total
D	101	99	200
$\bar{D}$	167	33	200

[a] $\psi = 4.96$.

$\psi(Dz|E) = 410.7$, $\psi(Ez|\bar{D}) = 9.86$ and $\psi(Ez|D) = 9.98$. These figures indicate that z is related to both exposure and disease, i.e., that is a confounding factor. Studying the relation of z to the disease could be more important than studying the present exposure and disease association, since the large values of $\psi(Dz|\bar{E})$ and $\psi(Ez|E)$ indicate that z is a predictor of the disease. It might be suspected that $\psi(z_1) = \psi(z_2) < \psi$ because the data were matched on a predictor of the disease. However, this is not true. Roughly, the strength of the influence of the confounding factor z upon cause-effect association can be measured by the absolute value of

$$\tau = [\psi(Dz|\bar{E}) - 1][\psi(Ez|\bar{D}) - 1]$$

If $\tau > 0$, then

$$\psi(z_1) = \psi(z_2) < \psi$$

and if $\tau < 0$, then

$$\psi(z_1) = \psi(z_2) > \psi$$

This suggests that when z is a predictor of the disease, it is not its role as predictor but rather its relation to the exposure that leads to an under-or overestimation problem. Thus how strongly the strength of association of z with disease status is not logically related to overmatching. In concluding this section we emphasize the necessity of checking whether a factor which is identified by our statistical methods as a confounding factor is an intermediate factor in the causal pathway before matching upon it.

Effect Modification

"z is related to exposure" is defined in the previous section by "at least one of $P(E|D,z)$ and $P(E|\bar{D},z)$ depends on z". It is not unnatural to suppose that the influence of $z = z_1$ on the exposure probability among cases is equal to that among controls for any fixed z, so that if $P(E|D,z)$ depends on z, $P(E|\bar{D},z)$ also depends on z, and vice versa. The principle of pairwise-matching (stratification), where a control with the same value of z as a case is selected for comparison seems to have been based upon this idea. Cox's model (*14*) to prove the optimality of the McNemar test for matched pairs data, Cornfield and Haenszel's discussion (*15*) of the relative risk estimator for matched pairs data, Gart's method (*16*) of calculating a summary statistic by estimating the common odds ratio by strata, and many other studies have all assumed it implictly or explicitly. However, this is not true in general. Miettinen (*17*) noted this fact and introduced effect modification for the purpose of explaining it. That effect modification is equivalent to second-order interaction is well known among statisticians.

Let $\psi(z_1)$ and $\psi(z_2)$ be the relative risks of disease associated with exposure in strata z_1 and z_2. If $\psi(z_1) \neq \psi(z_2)$, then we can say that the effect of exposure upon disease status in stratum z_1 is not equal to that in stratum z_2 (because of the existence of second-order interaction). Such a factor z has been called an effect modifier. The magnitude of effect modification is measured by either

$$\text{e.m.} = \psi(z_2)/\psi(z_1) \tag{12}$$

$$\text{e.m.} = \psi(Ez|D)/\psi(Ez|\bar{D}) \tag{13}$$

or

$$\text{e.m.} = \psi(Dz|E)/\psi(Dz|\bar{E}) \tag{14}$$

where $\psi(Ez|D)$, $\psi(Ez|\bar{D})$, and $\psi(Dz|\bar{E})$ are defined as in the previous section. Effect modification will be discussed further in the next section.

A Model with Two Risk Factors to Illustrate Confounding and Effect Modification

The following discussion regarding the joint effect of two risk factors in inducing disease should clarify understanding of confounding and effect modification.

Let A and B be factors suspected of inducing disease. Let us suppose for simplicity that both of them are dichotomous. Table 7 summarizes a prospective framework of probability distributions, where $P(D|\bar{A},\bar{B})$ is the probability of disease in an individual exposed to neither A nor B, $P(D|A,\bar{B})$ [$P(D|\bar{A},B)$] is the probability of disease after exposure to A (B) alone, and $P(D|A,B)$ is the disease probability after exposure to both A and B. Then $P(D|\bar{A},B) / P(D|\bar{A},\bar{B})$ is the relative risk due to B among those unexposed to A and $P(D|A,B) / P(D|A,\bar{B})$ is the relative risk due to B among those exposed to A. If these relative risks

Table 7.

	$\bar{A}$		A	
	$\bar{B}$	B	$\bar{B}$	B
D	$P(D\|\bar{A},\bar{B})$	$P(D\|\bar{A},B)$	$P(D\|A,\bar{B})$	$P(D\|A,B)$
$\bar{D}$	$1 - P(D\|\bar{A},\bar{B})$	$1 - P(D\|\bar{A},B)$	$1 - P(D\|A,\bar{B})$	$1 - P(D\|A,B)$

are equal, one might say that factors A and B have no joint effect (no interaction), then we might say there is a positive joint effect (positive interaction); we will say there is a negative joint effect (negative interaction) otherwise. Let us define γ, through Eq. (15):

$$\frac{P(D|A,B)}{P(D|\bar{A},\bar{B})} = \frac{P(D|A,\bar{B})}{P(D|\bar{A},\bar{B})}\,\frac{P(D|\bar{A},B)}{P(D|\bar{A},\bar{B})}\,\gamma' \qquad (15)$$

We will say there is no interaction if $\gamma' = 1$, positive interaction if $\gamma' > 1$, and negative interaction if $\gamma' < 1$. When $\gamma' = 1$, equation (1) states that the relative risk due to exposure to both A and B is the product of the relative risks due to exposure A and B separately. For this reason, (1) is frequently called a multiplicative risk model.

Setting

$$\begin{cases}\Delta_A = \log\,[P(D|A,\bar{B})/P(D|\bar{A},\bar{B})] \\ \Delta_B = \log\,[P(D|\bar{A},B)/P(D|\bar{A},\bar{B})] \\ \gamma = \log\,\gamma'\end{cases} \qquad (16)$$

Eq. (15) is equivalent to

$$\log\frac{P(D|A,B)}{P(D|\bar{A},\bar{B})} = \Delta_A + \Delta_B + \gamma \qquad (17)$$

where $\gamma = 0$, $\gamma > 0$ and $\gamma < 0$ indicate no interaction, positive interaction, and negative interaction, respectively.

Next, let us recast this example in a case-control framework. The data are presented in Table 8, where $P_i(\bar{A}\bar{B})$ is the probability of disease when there is no exposure to either A or B, P_i $(A\bar{B})$ $[P_i(\bar{A}B)]$ is the probability after exposure to A (B) alone, and $P_i(AB)$ is the probability after exposure to both A and B, among cases ($i = 1$) and controls ($i = 0$). Let us define

$$\begin{cases}\mu_A = \log\,[P_0(A\bar{B})/P_0(\bar{A}\bar{B})] \\ \mu_B = \log\,[P_0(\bar{A}B)/P_0(\bar{A}\bar{B})] \\ \alpha = \log\,[P_0(AB)/P_0(\bar{A}\bar{B})] - \mu_A - \mu_B\end{cases} \qquad (18)$$

Further, let us accept Cornfield's assumption that the prevalence of this disease is small enough so that the relative risks are approximated by the corresponding odds ratios. Then it follows from Eqs. (16)-(18) that

$$\begin{cases}\log\,[P_1(A\bar{B})/P_1(\bar{A}\bar{B})] = \mu_A + \Delta_A \\ \log\,[P_1(\bar{A}B)/P_1(\bar{A}\bar{B})] = \mu_B + \Delta_B \\ \log\,[P_1(AB)/P_1(\bar{A}\bar{B})] = \mu_A + \mu_B + \alpha + \Delta_A + \Delta_B + \gamma\end{cases} \qquad (19)$$

which is an extension of the well-known logistic linear model for 2 × 2 table analysis [see, for example, Cox (*18*)] to a 2 × 4 table. The multiplicative risk model (15) in a prospective framework is, therefore, equivalent to (18) and (19) in a case-control framework, under Cornfield's assumption. The parameters of interest are Δ_A, the log relative risk of A, Δ_B, the log relative risk of B, and their interaction γ. Thus parameters μ_A, μ_B, and α are nuisance parameters introduced by the case-control framework.

Finally, let us suppose that cases and controls have been stratified in the design by means of the factor A, i.e., unexposed and exposed to A. Then we have Table 9.

Table 8. Joint distribution of exposure to A and B in cases and controls.

	$\bar{A}$		A		
	$\bar{B}$	B	$\bar{B}$	B	Total
Cases	$P_1(\bar{A}\bar{B})$	$P_1(\bar{A}B)$	$P_1(A\bar{B})$	$P_1(AB)$	1
Controls	$P_0(\bar{A}\bar{B})$	$P_0(\bar{A}B)$	$P_0(A\bar{B})$	$P_0(AB)$	1

Table 9.

	$\bar{A}$			A		
	$\bar{B}$	B	Total	$\bar{B}$	B	Total
Cases	$1 - P_1(B\|\bar{A})$	$P_1(B\|\bar{A})$	1	$1 - P_1(B\|A)$	$P_1(B\|A)$	1
Controls	$1 - P_0(B\|\bar{A})$	$P_0(B\|\bar{A})$	1	$1 - P_0(B\|A)$	$P_0(B\|A)$	1

where P_i $(B|A)$ $[P_i(B|\bar{A})]$ is the probability of exposure to B in the stratum A $(\bar{A})$ for cases ($i = 1$) and controls ($i = 0$). It follows from Eqs. (18) and (19) that

$$\log\frac{P_i(B|\bar{A})}{1 - P_i(B|\bar{A})} = \mu_B + i\Delta_B$$

$$\log\frac{P_i(B|A)}{1 - P_i(B|A)} = \mu_B + \alpha + i(\Delta_B + \gamma) \qquad \text{for } i = 0, 1 \qquad (20)$$

Relative risks due to B within strata $\bar{A}$ and A are given by $\psi = \exp\{\Delta_B\}$ and $\psi = \exp\{\Delta_B + \gamma\}$," respectively.

Summarizing the above discussion, we may conclude that μ_A and Δ_A are deleted from model (19) when we stratify on factor A; in other words, as has been well known, we should not stratify (or match) cases and controls on a factor that is under investigation. α and γ may not be deleted from model (19) after stratification; in other words odds ratios for B within strata $\bar{A}$ and A are not equal, unless there is no interaction between A and B in the sense of relative risk. Since the factor A as considered in the

framework of model (20) is identical to the variable z discussed in the previous sections, we may say that Miettinen's effect modifier is a factor z that has some interaction with the factor under investigation. The discussion above regarding confounding variables is illustrated by model (20) as follows. Let us set $z = A$, $z_1 = \bar{A}$ and $z_2 = A$.

If $\Delta_z = \gamma = 0$, then z is not a confounding factor.

If $\gamma = 0$ and $\Delta_z > 0$, then

$\psi = \psi(z)$ for all z is equivalent to $\alpha = 0$;

$\psi > \psi(z_1) = \psi(z_2)$ if and only if $\alpha > 0$;

$\psi < \psi(z_1) = \psi(z_2)$ if and only if $\alpha < 0$.

Further, since $\alpha = 0$, $\alpha > 0$ and $\alpha < 0$ if and only if the joint distributions of exposure to B and z among the cases are independent, positively and negatively correlated respectively, we have:

$\psi = \psi(z_1) = \psi(z_2)$ if and only if the joint distribution of exposure to B and z in the cases are independent;

$\psi > \psi(z_1) = \psi(z_2)$ if and only if those of B and z in the cases are positively correlated;

and $\psi < \psi(z_1) = \psi(z_2)$ if and only if those of B and z in the cases are negatively correlated. In the first of these cases, z is not a confounding factor.

If $\gamma \neq 0$, then z is a confounding factor.

The strength of the influence of the confounding factor upon exposure and disease association may be measured by Eq. (21)

$$\tau = |(\exp\{\alpha\} - 1)(\exp\{\Delta_A\} - 1) + (\exp\{\gamma\} - 1)(1 + \exp\{\mu_A\})\exp\{\alpha\}\exp\{\Delta_A\}| \quad (21)$$

Many of the authors' studies cited above have assumed essentially that $\gamma = 0$. Note that the application of the maximum likelihood method to the model with $\gamma = 0$ provides the same summary odds ratio as Gart (*16*). However, if further risk factors were ignored in the study, $\gamma = 0$ still could not be expected even if A were definitely known not to induce disease, since the value of γ could be influenced by some ignored factor which had interaction with the factor under investigation. Further, suppose that both A and B are (strong) risk factors and have no synergistic relationship in inducing disease. Then γ should be negative since it measures interaction on the multiplicative scale, whereas a synergistic relationship is measured on the additive scale (3).

The model of Eq. (20) agrees with a special case of that considered by Prentice (2). He called z (i.e., factor A) a confounding factor if $\alpha \neq 0$. However, this may not be true. A counter example is given in Table 10. Here $\psi(z_1) = \psi(z_2) = \psi = 6$, so z is not a confounding factor, yet $\alpha = -0.85$.

The definition of a confounding factor is equivalent to the "collapsibility of categories" discussed by Bishop, Feinberg, and Holland (*19*). The results summarized above agree with their deductions, which were obtained by means of log linear models.

Table 10. Comparison of pooled and unpooled data.

	Unpooled data[a]					Pooled data[b]		
	z_1		z_2					
	$\bar{B}$	B	$\bar{B}$	B	Total	$\bar{B}$	B	Total
Cases	5	15	35	45	100	40	60	100
Controls	10	5	70	15	100	80	20	100

[a] $\psi(z_1) = 6$, $\psi(z_2) = 6$.
[b] $\psi = 6$.

$\alpha = \log(15/10) - \log(5/10) - \log(70/10) = -0.85$

Classification and Stratification

In the model of Eq. (19) controls are selected from a population comparable to the population of cases; then it is determined into which of the classes $\bar{A}\bar{B}$, $\bar{A}B$, $A\bar{B}$ or AB they fall. On the other hand, in the model of Eq. (20), a predetermined number of controls are selected among those individuals who have $\bar{A}$ and A, respectively, and they are then classified according to whether they have $\bar{B}$ or B. Therefore, we could say that the first model is based on classification, whereas the second model is based on stratification. The difference lies in the sampling strategies. The first model provides a relative risk, not only for factor B but also for factor A. Even though A is thought not to induce disease, we may find the relative risk greater than 1. Investigation of the reason could often provide further information. For example, place of residence is normally not a risk factor for lung cancer, yet we might find the relative risk for some location greater than 1. Investigation could reveal the presence of certain suspect industries in the region. Or perhaps we will find a relative risk for A of 1 but with γ greater (smaller) than zero. Such a finding would be especially interesting, since it would suggest that factor A alone is not the risk factor, but that it amplifies (diminishes) the relative risk of B if it operates together with B. A significant advantage of the classification model is its flexibility. It permits us to identify and to evaluate the influence of confounding factors. It also provides estimates of relative risks free from the influence of these factors. Further, as will be seen in a subsequent section, it also provides estimates of relative risks adjusted for combinations of factors. Generally, the model (19) provides more information than the model (20).

A drawback of the sampling strategy which leads

to model (19) is that the estimates of Δ_B and γ are likely to be influenced by any bias present in the selection of controls. This should be seriously considered in a case-control study, since it further complicates the usual difficulties in selecting controls.

Another advantage of stratification is that we can increase our precision in estimating Δ_B and γ by selecting an appropriate number of controls from each stratum.

Summarizing the above discussion, we recommend the following strategy: (1) stratify cases and controls by means of confounding variables which are definitely known not to induce disease and which are not of interest to the investigation; (2) classify cases and controls by means of confounding variables whose role in the induction of disease is known or suspected. An analytic model for this approach will be discussed in the next section.

Table 8, where cases and controls are classified by means of $\bar{A}\bar{B}$, $\bar{A}B$, $A\bar{B}$, and AB, can be broken into two 2×2 tables and analyzed. A beautiful analysis based on this approach was given by Prentice (*2*). His approach enables us to decrease the number of parameters to be estimated. However, the bias and precision of the estimated parameters are the same, at least asymptotically, in this model as in the model of Eq. (19). Because of this, and for the reasons mentigned above, analysis based on the model of Eq. (19) is recommended. If necessary, the relative risk within strata $\bar{A}$ and Acould be represented by exp $\{\Delta_B\}$ and exp$\{\Delta_B + \gamma\}$, which are sometimes called the relative risks for B adjusted for the factor A. Our method yields an estimate of summary relative risk when γ is set to zero.

A weak point of the analysis based on the model of Eq. (19) is when the number of cases and controls is small, since the usual methods for estimation of parameters employ asymptotic approximations. In this case Breslow's (20) recent approach is useful. He has given an exact analysis, considering all the marginal totals of the two 2×2 tables in Table 8 to be fixed. The model which he applied is the linear model for the log odds ratio, which is derived from our model, Eq. 20, as follows:

$$\log \frac{P_1(B|\bar{A})\,[1 - P_0(B|\bar{A})]}{[1 - P_1(B|\bar{A})]\,P_0(B|\bar{A})} = \Delta_B$$

$$\log \frac{P_1(B|A)\,[1 - P_0(B|A)]}{[1 - P_1(B|A)]\,P_0(B|A)} = \Delta_B + \gamma \qquad (22)$$

He made use of a computer program to carry out the exact analysis. However, as the number of cases and controls becomes large, computation time becomes prohibitive.

A Model Taking into Account Classification and Stratification Simultaneously, Where One Factor Assumes More Than Two Values

Let us consider a model with simultaneous stratification and classification, where one variable can take on more than two values. We shall consider first a situation where there are two factors A and B. Let us suppose B is dichotomous, where A is trichotomous, with possible values A_0, A_1, A_2. Table 11 summarizes the probability distributions for cases and controls, where both are classified on A and B.

An analytic model for Table 11 is given in Eqs. (23)-(27),

$$\log (P_{i01}/P_{i00}) = \mu_B + i\Delta_B \qquad (23)$$

$$\log (P_{i10}/P_{i00}) = \mu_{A(1)} + i\Delta_{A(1)} \qquad (24)$$

$$\log (P_{i11}/P_{i00}) = \mu_{A(1)} + \mu_B + \alpha_{A(1)B} + i(\Delta_{A(1)} + \Delta_B + \gamma_{A(1)B}) \qquad (25)$$

$$\log (P_{i20}/P_{i00}) = \mu_{A(2)} + i\Delta_{A(2)} \qquad (26)$$

$$\log (P_{i21}/P_{i00}) = \mu_{A(2)} + \mu_B + \alpha_{A(2)B} + i(\Delta_{A(2)} + \Delta_B + \gamma_{A(2)B}) \qquad (27)$$

where $\Delta_{A(1)}$, $\Delta_{A(2)}$ and Δ_B are log odds ratios for the factors A_1, A_2, and B, $\gamma_{A(i)B}$ is the interaction of factors A_i and B ($i = 1, 2$), and the other parameters are nuisance parameters introduced by the case-control framework.

Next, let us expand on Table 11 by stratifying on certain confounding variables z and w, such as age and sex. Let us denote by $P_{ijk}(z,w)$ the probability P_{ijk} in the stratum specified by z and w. Then the analytic model is given by Eqs. (28)-(30).

$$\log [P_{i01}(z,w)/P_{i00}(z,w)] = \mu_B + z\alpha_{zB} + w\alpha_{WB} + zw\beta_{zwB} + i[\Delta_B + z\gamma_{zB} + w\gamma_{wB} + zw\delta_{zwB}] \qquad (28)$$

$$\log [P_{ij0}(z,w)/P_{i00}(z,w)] = \mu_{A(j)} + z\alpha_{zA(j)} + w\alpha_{WA(j)} + zw\beta_{zwA(j)} + i[\Delta_{A(j)} + z\gamma_{zA(j)} + w\gamma_{wA(j)} + zw\delta_{zwA(j)}] \qquad (29)$$

Table 11. Probability distributions of cases and controls.

	A_0		A_1		A_2		
	$\bar{B}$	B	$\bar{B}$	B	$\bar{B}$	B	Total
Cases	P_{100}	P_{101}	P_{110}	P_{111}	P_{120}	P_{121}	1
Controls	P_{000}	P_{001}	P_{010}	P_{011}	P_{020}	P_{021}	1

$$\log [P_{ijl}(z,w)/Pb_{i00}(z,w)] = \mu_{A(j)} + \mu_B + \alpha_{A(j)B} + z(\alpha_{zB} + \alpha_{zA(j)} + \beta_{zA(j)B}) + w(\alpha_{wB} + \alpha_{wA(j)} + \beta_{wA(j)B}) + zw(\beta_{zwB} + \beta_{zwA(j)} + \beta_{zwA(j)B}) + i[\Delta_{A(j)} + \Delta_B + \Delta_{A(j)B} + z(\gamma_{zB} + \gamma_{zA(j)} + \delta_{zA(j)B}) + w(\gamma_{wB} + \gamma_{wA(j)} + \delta_{wA(j)B}) + zw(\delta_{zwB} + \delta_{zwA(j)} + \delta_{zwA(j)B})] \quad (30)$$

for $i = 0, 1$ and $j = 1, 2$, where parameters $\gamma_{zA(j)}$, γ_{zB}, $\gamma_{wA(j)}$, and γ_{wB} are interactions of z and A_j, z and B, w and A_j, and w and B; $\delta_{zwA(j)}$, δ_{zwB}, $\delta_{zA(j)B}$ and $\delta_{wA(j)B}$ are the interactions of z, w, and A_j; z, w, and B; z, A_j, and B; and w, A_j, and B. $\delta_{zwA(j)B}$ is the interaction of z, w, A_j, and B; other parameters newly introduced are nuisance parameters introduced by the case-control framework.

Normally it would be rare to have information beyond third-order interactions. In the simpler case, similar results to those based on the above model could be obtained by applying a model which ignores the β and δ parameters in the above model.

Parameters for these models can be estimated by the weighted least squares method of Grizzle, Starmer, and Koch (*21*), or by the method of maximum likelihood intensively discussed in the book of Bishop, Feinberg, and Holland (*19*). The number of parameters in these models looks excessive. But I suggest that it is better to start from a saturated model and to undertake an iterative process to reach the most appropriate and simplest model that could explain the structure of data in detail; starting from the above model, first estimate all parameters, then examine them, deleting those that do not contribute significantly and finally develop a simplified model. The approach would be especially useful if a case-control tudy were an exploratory one intended to locate causal factors. If it is a confirmatory study, then we should use, of course, all information obtained from previous studies as well as existing knowledge to establish a simpler model for the initial model. Statistical methods, such as the "Akaike information criterion" (AIC) (22), all possible regressions (*23*), stepwise regressions (*24*), etc., can be applied to determine how many parameters should be included in the model. In my experience, the method employed by Grizzle, Starmer, and Koch is the most handy and efficient among others for that purpose, although special care is necessary in applying the method if empty cells exist.

Number of Cases and Controls

Generally, the number of confounding factors and the number of levels of each factor to be considered in the study are determined, therefore, based on the number of cases. If the group of cases is not large, then we must ignore some confounding factors or decrease the number of levels of certain factors, e.g., by collapsing the age categories into wider ranges for each stratum. If this process is suspected of introducing serious bias, we may have to switch to pair-matching. However, a well-known difficulty of matched pairs design lies in the selection of controls. Cochran (5) has estimated that the reservoir from which controls are to be selected must be at least six times the size of the number of cases. Prentice (7) proposed a method to liberalize the study design substantially and increase the estimating efficiency. This is a method of adjusting for the unavailability of a corresponding matched individual statistically in the analysis. The model proposed in the last section has the same property as Prentice's, when individuals are matched on z and w.

Special attention must be paid to the empty cells before collapsing the exposure categories or otherwise changing procedures in order to eliminate them, since they are likely to provide considerable information; for example, if the exposure categories are ordered in some way and there is a strong dose-response relationship with respect to that ordering, then extreme cells for the controls could well be empty. If such is the case the number of controls should be increased to eliminate the empty cells; if no such dose-response relationship is seen, then reliance on the previously discussed stratification on a selected set of confounding variables would be suitable. An advantage of the models discussed in previous sections is that even if, say, 10% of the cells for cases and controls are empty, we can use the information obtained from the 90% of the cells that are not empty to estimate parameters which will represent the structure of the data satisfactorily.

It is not yet well established how to determine how many cases are necessary when several confounding factors are taken into account. It depends both on financial restrictions and on the purpose of the study. Let us ignore the former and consider only the latter. Let A and B be suspected (dichotomous) risk factors which are of interest. If A is the target factor, then the familiar method discussed intensively in the book of Fleiss (*9*) may be applied to a 2×2 table, obtained by ignoring the factor B, to get a rough estimate of the required number of cases. If A and B are equally important factors and the investigation is intended to determine the effects of both A and B, as well as their interactions, in inducing disease, then a test of the degree of interaction could help to determine the required number of cases. If there is a priori evidence that interaction does not exist, a rough estimate of the required number could be obtained by applying the above method to two 2×2 tables, one obtained

by ignoring the factor B and the other by ignoring factor A, and by using the larger number.

An Illustrative Example of the Method

Information on lifetime smoking and occupational histories for 101 white male coastal Georgia residents diagnosed with lung cancer during 1970-76, and for 203 white male age- and residence-matched hospital controls diagnosed with conditions other than lung cancer or lung disease, was obtained by personal interview.* Each case and control was classified into one of three smoking levels based on his cigarette smoking history: (1) none or light (< ½ pack/day) (includes individuals who quit smoking at least 10 years before diagnosis); (2) moderate (½ to 1½ packs/day); (3) heavy (2 or more packs/day). Each individual was also categorized (yes/no) as to whether he had ever been employed in each of the shipbuilding or construction industries. The resulting responses are listed in Table 12. A model with 22 parameters, similar to the one discussed in the last section, was set up as a preliminary model. A stepwise procedure was carried out, using the weighted least-squares method of Grizzle, Starmer, and Koch (21), and 10 of the 22 parameters were eliminated, leaving the model (31) as the one best reflecting the structure of the data given in Table 12.

$$\log(P_{0ijl}/P_{0000}) = (2-i)i\mu_{A(1)} + [i(i-1)/2]\mu_{A(2)} + j\mu_B + l\mu_C + i(2-i)l\alpha_{A(1)C} + [i(i-1)l/2]\alpha_{A(2)C} + [i(i-1)j/2]\alpha_{A(2)B}$$

$$\log(P_{1ijl}/P_{1000}) = \log(P_{0ijl}/P_{0000}) + (2-i)i\Delta_{A(1)} + [i(i-1)/2]\Delta_{A(2)} + j\Delta_B + l\Delta_C + [i(3-i)\,l/2]\gamma_{AC} \qquad i = 1, 2; j = 0, 1; l = 0, 1 \qquad (31)$$

where $\Delta_{A(1)}$ and $\Delta_{A(2)}$ are the log relative risks due to moderate and heavy smoking, respectively, as compared to none or light; Δ_B the log relative risk due to employment in the shipbuilding industry as compared to nonemployment in the industry; and Δ_C the log relative risk due to employment in the construction industry as compared to non-employment in the industry; interactions of A_1 and C and A_2 and C are assumed to be equal and are represented by γ_{AC}; the other parameters are nuisance parameters introduced by the case-control framework.

Table 13 summarizes estimates of the 12 parameters in of the model (31) and their standard deviations. Computed values for the variances and covariances of estimates of $\Delta_{A(1)}$, $\Delta_{A(2)}$, Δ_B, and Δ_C, and γ_{AC} are summarized in Table 14.

The influence of an empty cell in Table 12 was

*The data presented here, which were provided to the author by Dr. William J. Blot, represent only a part of a complete case-control study; they were selected for illustrative purposes and should not be used to draw inferences about cancer risk. A detailed description of the Georgia study and a full report of the results is given elsewhere (25).

Table 13. Estimated values of parameters used in the model shown in Eq. (31) and their standard deviations (SD).[a]

Parameters	Estimates	SD.
$\mu_{A(1)}$	−0.6070	0.1887
$\mu_{A(2)}$	−1.2040	0.2393
μ_B	−1.6759	0.2036
μ_C	−1.0616	0.2236
$\alpha_{A(1)C}$	0.7157	0.3180
$\alpha_{A(2)C}$	−0.2510	0.4289
$\alpha_{A(2)B}$	−1.4812	0.6179
$\Delta_{A(1)}$	2.4214	0.5130
$\Delta_{A(2)}$	3.0165	0.5385
Δ_B	0.6559	0.3349
Δ_C	1.5488	0.6142
γ_{AC}	−1.3960	0.6882

[a]Relative risks: smoking, exp $\{\Delta_{A(1)}\}$ = 11.26 (moderate), exp $\{\Delta_{A(2)}\}$ = 20.42 (heavy); ship building, exp $\{\Delta_B\}$ = 1.93; construction, exp $\{\Delta_C\}$ = 4.71.

Table 12. Exposure to shipbuilding, construction, and smoking for lung cancer cases and controls.[a]

	A_0				A_1				A_2				
	B_0		B_1		B_0		B_1		B_0		B_1		
	C_0	C_1	C_0	C_1	C_0	C_1	C_0	C_1	C_0	C_1	C_0	C_1	Total
Cases	4	5	1	3	25	17	6	8	23	7	1	1	101
	3.7	6.0	1.3	2.2	22.6	18.9	8.1	6.7	22.5	7.0	1.9	.6	101
Controls	68	22	10	5	37	23	5	7	20	5	1	0	203
	65.5	22.7	12.2	4.2	35.7	25.2	6.7	4.7	19.7	5.2	.8	.2	203

[a]Predictions from model of Eq. (31). Factors: A, smoking (A_0 = non smoking or light, A_1 = moderate, A_2 = heavy); B, ship building (B_0 = unexposed, B_1 = exposed); C, construction (C_0 = unexposed, C_1 = exposed).

Table 14. Computed variances and covariances of the estimates of $\Delta_{A(1)}$, $\Delta_{A(2)}$, Δ_B, Δ_C, and γ_{AC}.

	$\Delta_{A(1)}$	$\Delta_{A(2)}$	Δ_B	Δ_C	γ_{AC}
$\Delta_{A(1)}$	0.2631				
$\Delta_{A(2)}$	0.2220	0.2900			
Δ_B	−0.0005	0.0165	0.1122		
Δ_C	0.2129	0.2104	−0.0148	0.3772	
γ_{AC}	−0.2556	−0.2333	0.0041	−0.3758	.4740

determined to be negligible — replacing the zeroes with values smaller than 1/6 had almost no effect on the computation. To check the validity of the model, the number of cases and controls in each category was predicted from Eq. (31) by using estimated values for the parameters. The predicted values, summarized in the second and fourth rows of Table 12, agree fairly well with the original data. Therefore, the model appears to describe the structure of the data nicely.

Suppose that our primary interest is in the association between exposure to A and lung cancer, with B and C as additional factors. Adjusted relative risks due to exposure to A, adjusted for B and C, have a structure represented in Table 15. Estimates of these relative risks are obtained by substituting the values for the parameters shown in Table 13. Table 15 shows that the relative risks within stratum B_0 are equal to those within stratum B_1. This results because in the model in [Eqs. (31)] $\gamma_{AiB} = 0$, $i = 1, 2$; i.e., B is not an effect modifier. B is, however, a confounding factor, since $\alpha_{A(2)B} \neq 0$. The influence of B on the association of A and lung cancer is estimated by

$$\tau = (\exp\{\Delta_B\} - 1)(\exp\{\alpha_{A(2)B}\} - 1) = -0.72$$

This indicates a slight underestimate of the relative risk if B is ignored. On the other hand, C is an effect modifier, since $\gamma_{AC} \neq 0$. The magnitude of the effect modification is estimated by e.m. = $\exp\{\gamma_{AC}\}$ = 0.25. Thus the relative risk due to A among those

Table 15. Structure of adjusted log relative risks of exposure to smoking (A) adjusted for exposure to shipbuilding (B) and construction (C).

Study factor	Adjustment factors			
	B_0		B_1	
	C_0	C_1	C_0	C_1
A_0	1	1	1	1
A_1	$\Delta_{A(1)}$	$\Delta_{A(1)}$	$\Delta_{A(1)}$	$\Delta_{A(1)} + \gamma_{AC}$
A_2	$\Delta_{A(2)}$	$\Delta_{A(2)} + \gamma_{AC}$	$\Delta_{A(2)}$	$\Delta_{A(2)} + \gamma_{AC}$

Table 16. Structure of adjusted log relative risks of exposure to shipbuilding (B) adjusted for exposure to smoking (A) and construction (C).

Study factor	Adjustment factor					
	A_0		A_1		A_2	
	C_0	C_1	C_0	C_1	C_0	C_1
B_0	1	1	1	1	1	1
B_1	Δ_B	Δ_B	Δ_B	Δ_B	Δ_B	Δ_B

exposed to C is modified to a quarter of that among those unexposed to C. The smaller value of the relative risk due to A among those exposed to C occurs for the reason discussed above. The influence of C on the association is estimated by $\max(|\tau_{A(1)}|, |\tau_{A(2)}|)$, where

$$\begin{aligned}\tau_{A(1)} &= (\exp\{\Delta_c\} - 1)(\exp\{\alpha_{A(1)C}\} - 1) \\ &\quad + (\exp\{\gamma_{AC}\} - 1)(1 + \exp\{\mu_{A(1)}\})\exp\{\alpha_{A(1)C}\} \\ &\quad \exp\{\Delta_{A(1)}\} = -22.90 \\ \tau_{A(2)} &= -16.36\end{aligned} \tag{32}$$

If the Mantel-Haenszel method were applied in the analysis, we would have to ignore either B or C, or to pool A_1 and A_2, since there are several cells in Table 12 whose entries are quite small. Because A is our study factor, we would prefer no pooling of A. In that case, B should be ignored, since its τ value is quite close to zero compared to the corresponding value for C.

Next, let us assume that B is our study factor, and A and C additional factors. Adjusted relative risks due to exposure to B, adjusted for A and C, have the structure represented in Table 16. All the entires in a row are equal. This results because A and C are not effect modifiers, i.e., in Eq. (31) $\gamma_{A(i)B} = \gamma_{BC} = 0$, $i = 1, 2$. Since $\alpha_{A(2)B} \neq 0$, A is a confounding factor whose influence on the association is estimated by $\max(|\tau_1|, |\tau_2|)$, where

$$\begin{aligned}\tau_{A(1)} &= (\exp\{\Delta_{A(1)}\} - 1)(\exp\{\alpha_{A(1)B}\} - 1) = 0 \\ \tau_{A(2)} &= (\exp\{\Delta_{A(2)}\} - 1)(\exp\{\alpha_{A(2)B}\} - 1) = -15.00\end{aligned} \tag{33}$$

The fact that $\tau_{A(2)}$ is negative indicates than an underestimate of the relative risk will result if A is ignored in the study. Since $\alpha_{BC} = 0$, C is not a confounding factor in the assocation of exposure to B and lung cancer. Thus, if the Mantel-Haenszel method were applied in the analysis, C should be ignored for two reasons: (1) C is related to disease but unrelated to exposure; (2) the problem of small cell entries discussed above. When C is ignored and the Mantel-Haenszel method is applied, the summary statistic $\psi = 1.87$, which is fairly close to $\exp\{\Delta_B\} = 1.93$.

Conclusion

Identification of confounding factors, evaluation of their influence on exposure and disease association, and the introduction of proper devices, such as matching, stratification, classification and others, into the design to block the influence of these factors are very important in designing case-control studies. We presented a theoretical review of recent developments in this area, based on a statistical definition of confounding factor. With such a definition, medical knowledge is required to determine whether or not confounding factors identified by our methods are intermediate factors in the casual pathway between the study factor and the disease. If a confounding factor is an intermediate factor we should not match on it (overmatching); if not, we must introduce some device to block its influence. Stratification, or matched pairs design in its extreme form, have been the main design devices for blocking the influence of confounding factors. However, the identification of a confounding factor and the evaluation of the strength of its influence on the association are not feasible from data selected by such sampling strategies. However, identification and evaluation can be achieved through a random sampling of cases and controls from a population and their classification into categories, based on known and suspected confounding factors. This paper suggested stratification of cases and controls on those confounding variables which are definitely known not to induce disease and which are not of interest in the study, and classification of cases and controls on confounding variables which are known or suspected of inducing disease. Logistic linear models were introduced for the combined purpose of identification of confounding factors, evaluation of their influence on the relative risk, and analysis of the data. They are extensions of the well-known logistic model for 2×2 table analysis, as applied in a case-control study by Prentice (*2*). The paper recommends starting from such a model, then following an iterative process to derive the most appropriate and simplest model that will explain the structure of the data in detail. If the study is a preliminary one, the resulting model can be used to identify the confounding factors and evaluate the strength of their influence on the cause-effect association in preparation for a follow-up study. Estimates of relative risks and interactions are also obtained. Estimates of adjusted relative risks, adjusted for combinations of factors, are also obtained by simple manipulation of the estimated relative risks from the model. In contrast to the method of Prentice (*2*), this approach requires only a single computer calculation, not successive iterations, but it shares with Prentice (*2*) the ability to adjust for the unavailability of matches for some individuals, if pair-matching is applied to certain confounding factors such as age. Thus it can substantially liberalize the study design and increase estimating efficiency.

Bishop, Feinberg, and Holland (*19*) have discussed thoroughly the analysis of frequency data by log linear models. The definition of a confounding factor given in this paper is identical to their concept of "collapsibility of categories." Thus their general approach could be used quite effectively in case-control studies. Statisticians may prefer their approach. However, it could result in useless statistical manipulation for epidemiologists unless statisticians understand precisely traditional epidemiological ideas which have been developed in the field. We hope that discussions in the present paper will help them to understand such ideas and apply them in their epidemiological research.

Case-control studies are becoming more complex in design and analysis, where, as was pointed out by McKinlay (*2*), "emphasis is increasingly being given to the investigation and estimation of multivariate sources of variation rather than simply being restricted to the removal of bias from a single comparison." Although the design and analysis of case-control studies using logistic linear models as introduced in the present paper seem complicated, such models, as well as the log linear model discussed by Bishop, Feinberg, and Holland (19), will play a central role in such studies.

The main part of the present study was carried out while I was a Visiting Scientist at the Environmental Epidemiology Branch, National Cancer Institute, Bethesda, Md., 20014, U.S.A. I am grateful to Dr. Joseph F. Fraumeni, Chief of the Branch, and to Dr. William J. Blot, my sponsor, who provided me with an excellent opportunity for examining statistical problems in cancer epidemiology.

REFERENCES

1. Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22: 719 (1959).
2. Prentice, R. Use of the logistic model in retrospective studies. Biometrics 32:599 (1976).
3. MacMahon, B., and Pugh, T. F. Epidemiology: Principles and Methods. Little, Brown & Co., Boston, 1970.
4. McKinlay, S. M. The design and analysis of the observational study — a review. J. Am. Statist. Assoc. 70: 503 (1975).
5. Cochran, W. G. The planning of observational studies of human populations. J. Roy. Statist. Soc. A198 (Part 2): 234 (1965).
6. Lilienfeld, A. M. Foundations of Epidemiology. Oxford Univ. Press, New York, 1976.
7. Cornfield, J. A method of estimating comparative rates from clinical data. Application to cancer of lung, breast and cervix. J. Natl. Cancer Inst. 11: 1269 (1951).

8. Berkson, J. Smoking and lung cancer: Some observations on two recent reports. J. Am. Statist. Assoc. 53: 28 (1958).
9. Fleiss, J. L. Statistical Methods for Rates and Proportions. Wiley, New York, 1973.
10. Worcester, J. Matched samples in epidemiologic studies. Biometrics 20: 840 (1964).
11. Miettinen, O. S. Matching and design efficiency in retrospective studies. Am. J. Epidemiol. 91: 111 (1970).
12. Hardy, R. J., and White, C. Matching in retrospective studies. Am. J. Epidemiol. 93: 75 (1971).
13. Fisher, L., and Patil, L. Matching and unrelatedness. Am. J. Epidemiol. 100: 347 (1974).
14. Cox, D. R. Two further applications of a model for binary regression. Biometrika 45: 562 (1958).
15. Cornfield, J., and Haenszel, W. Some aspects of retrospective studies. J. Chronic Dis. 11: 523 (1960).
16. Gart, J. J. On the combination of relative risk. Biometrics 18: 601 (1962).
17. Miettinen, O. S. Confounding and effect modification. Am. J. Epidemiol. 100: 350 (1974).
18. Cox, D. R. Analysis of Binary Data. Methuen, London, 1969.
19. Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, Mass., 1975.
20. Breslow, N. Regression analysis of the log odds ratio: a method for retrospective studies. Biometrics 32: 409 (1976).
21. Grizzle, J. E., Starmer, C. F., and Koch, G. G. Analysis of categorical data by linear models. Biometrics 25: 489 (1969).
22. Akaike, H. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6): 716 (1974).
23. Ito, P. K., and Kudo, A. Some logical issues in interpreting multivariate data by means of regression analysis. In: Proceedings of the 8th International Biometric Conference, L. C. A. Corsten and T. Postelunicu, Eds., Vol. 85, 1975.
24. Draper, N. R., and Smith, H. Applied Regression Analysis. Wiley, New York, 1966, Chapt. 6.
25. Blot, E. J., Harrington, J. M., Toredo, A., Hoover, R., Heath, C. W., and Fraumeni, J. F. Lung cancer after employment in shipyards during World War II. N. Engl. J. Med. 299: 620 (1978).